Arthritis Care & Research
Vol. 74, No. 7, July 2022, pp 1179–1187
DOI 10.1002/acr.24563

AMERICAN COLLEGE of RHEUMATOLOGY
Empowering Rheumatology Professionals

Kellgren/Lawrence Grading in Cohort Studies: Methodological Update and Implications Illustrated Using Data From a Dutch Hip and Knee Cohort

Erin M. Macri, Jos Runhaar, Jurgen Damen, Edwin H. G. Oei, and Sita M. A. Bierma-Zeinstra

Objective. The Cohort Hip and Cohort Knee (CHECK) is a cohort of middle-aged individuals with hip or knee pain. Radiographs were assigned Kellgren/Lawrence (K/L) scores under different conditions at each follow-up visit for 10 years. We aimed to describe and consolidate each scoring approach, then illustrate implications of their use by comparing baseline K/L scores assigned using 2 of these approaches, and evaluating their respective associations with joint replacement and incident radiographic osteoarthritis (ROA).

Methods. We compared baseline K/L scores assigned to hips and knees using 2 scoring approaches: 1) assigned by senior researchers to baseline images alone and 2) assigned by trained readers, with images read paired and in known sequence with up to 10 years of follow-up radiographs (Poisson regression). We evaluated the associations of baseline ROA (any: K/L grade ≥1; established: K/L ≥2) with joint replacement, and of K/L 1 joints with incident established ROA (survival analysis).

Results. Of 1,002 participants (79% women, mean ± SD age 55.9 ± 5.2 years, body mass index 26.2 ± 4.0 kg/m^2), the second scoring approach had 2.4 times (95% confidence interval [95% CI] 1.8–3.1 for knees) and 2.9 times (95% CI 2.3–3.7 for hips) higher prevalence of established ROA than the first approach. Established hip ROA had a higher risk of joint replacement using the first approach (hazard ratio [HR] 24.2 [95% CI 15.0–39.8] versus second approach HR 7.7 [95% CI 4.9–12.1]), as did knees (HR 19.3 [95% CI 10.3–36.1] versus second approach HR 4.8 [95% CI 2.4–9.6]). The risk of incident ROA did not differ by approach.

Conclusion. This study demonstrates that evaluating ROA prevalence and predicting outcomes depends on the scoring approach.

INTRODUCTION

The presence and severity of knee or hip radiographic osteoarthritis (ROA) is commonly graded using the Kellgren/Lawrence (K/L) method (1). This semiquantitative approach primarily evaluates osteophytes and joint space narrowing to assign a score between 0 (no ROA) to 4 (severe ROA) (1–3). ROA is typically defined as K/L grade ≥2.

In cohort studies, standardized procedures to assign K/L scores include using a grading atlas, blinding readers to clinical features (e.g., pain), and reading radiographs paired with known sequence order (4–11). Reading single images (blinded to identity and sequence) is less sensitive to ROA progression compared to reading paired images, regardless of whether sequence is known (4,5,7). Reading paired images with known sequence has higher interrater reliability and sensitivity to ROA progression (4–11). However, blinding to sequence reduces bias (7), so although both methods do not significantly differ (5), some cohorts blind readers to sequence (12).

Reading conditions like single image versus paired are a source of error that can lead to misclassifying individuals regarding ROA prevalence (4–11). Other factors also influence scores, notably the somewhat arbitrary and subjective distinction between K/L grades 1 and 2 (13). Image-related factors include image acquisition plane, radioanatomic positioning, and image quality (14,15). Reader-related

The Cohort Hip and Cohort Knee study was initiated and funded by the Dutch Arthritis Association. Dr. Macri's work was supported by a Canadian Institutes of Health Research Banting Postdoctoral Fellowship. Dr. Bierma-Zeinstra's work was supported by The Netherlands Organization for Health Research and Development, Stichting Het Centraal Ziekenfonds, European Union, Foreum, Dutch Arthritis Association, and the Osteoarthritis Research Society International.

Erin M. Macri, PhD, Jos Runhaar, PhD, Jurgen Damen, MD, GP, PhD, Edwin H. G. Oei, MD, PhD, Sita M. A. Bierma-Zeinstra, PhD: Erasmus University Medical Center Rotterdam, Rotterdam, The Netherlands.

Dr. Bierma-Zeinstra has received consulting fees from Pfizer (less than $10,000). No other disclosures relevant to this article were reported.

Address correspondence to Erin M. Macri, PhD, Doctor Molewaterplein 40, 3015 GD Rotterdam, The Netherlands. Email: e.macri@erasmusmc.nl.

Submitted for publication June 20, 2020; accepted in revised form January 12, 2021.

SIGNIFICANCE & INNOVATIONS

- The prevalence of established hip or knee radiographic osteoarthritis (OA; Kellgren/Lawrence grade ≥2) was 2.4 to 2.9 times higher when assigning scores based on paired readings with known sequence as read by expert or trained readers compared to expert readers reading a single image.
- The highest hazard ratio for undergoing future hip or knee replacement in participants with established radiographic OA was when scores were read at a single time by expert readers (compared to paired reading in known sequence as read by expert or trained readers).
- The highest number of joints correctly classified as undergoing future hip or knee replacement occurred when images were read paired and in known sequence by expert or trained readers, and when OA was defined at a lower threshold of Kellgren/Lawrence grade ≥1.
- These findings highlight the importance of considering both radiographic scoring conditions as well as the threshold for defining OA when interpreting study results or designing new trials.

factors include training and experience (16,17). One cohort study reported "wobbles" over time whereby scores fluctuated between being classified as ROA or not (12). Further complicating the challenges of correctly classifying ROA, some researchers define ROA as K/L grade ≥1 (doubtful osteophytes), particularly in early OA research (18,19). Appreciating the extent to which reading conditions and ROA definitions influence misclassification could improve interpretation of study results and inform future study design.

The Cohort Hip and Cohort Knee (CHECK) study followed middle-aged individuals with knee or hip pain for 10 years (20). At each visit, radiographs were read and scored under different conditions as new images became available. Therefore, different CHECK publications (21–24) may have used different K/L scores, reflecting score wobble over time (12). This variation could confuse study interpretation among CHECK studies.

Our main aim was to describe and consolidate the knee and hip radiographic K/L scoring methods used in the CHECK cohort at each visit. Second, we aimed to compare the relative prevalence of baseline ROA using 2 different scoring approaches (single reading by expert readers versus paired readings of known sequence by expert and trained readers) and 2 definitions of ROA (K/L grade ≥1, K/L grade ≥2). Finally, we explored the association of baseline radiographic scores to 2 key outcomes: joint replacement and incident ROA.

PATIENTS AND METHODS

CHECK cohort. CHECK is a prospective multicenter cohort study (n = 1,002) (20). Recruitment took place at 10 hospitals throughout The Netherlands between 2002 and 2005. Individuals were ages 45–65 years with knee or hip symptoms for which they had not yet sought medical care, or who had first visited a general practitioner (GP) no more than 6 months prior to enrollment. Individuals were excluded if they had medical conditions that might otherwise explain their symptoms (e.g., rheumatic conditions, previous joint replacement); comorbidities preventing evaluations over 10 years; or malignancy in the previous 5 years. Ethics approval was provided by all participating centers, and participants provided informed written consent. Research adhered to the Helsinki Declaration.

Radiography. Radiographs of both knees and both hips were acquired at 5 time points (baseline, 2, 5, 8, and 10 years), unless a participant missed an appointment or withdrew from the study. Detailed protocols were followed at all study centers to ensure precise radioanatomic positioning, with the use of small metal balls, plexiglass frames, and foot-maps to ensure accurate, reproducible positioning across visits. We describe here only the acquired views needed for K/L scoring. For the knee, posteroanterior radiographs were taken with participants positioned in semiflexed weightbearing (23). For the hip, anteroposterior radiographs were taken with participants positioned in weightbearing.

K/L scoring procedures. Baseline images were first scored by a member of the CHECK steering committee. The steering committee consisted of senior investigators with substantial expertise in ROA research: 3 rheumatologists, 2 physical therapists, 1 rehabilitation physician, 1 physician, and 1 biologist. Prior to scoring images, the steering committee met to standardize scoring procedures, based on the original K/L scoring description (1–3), using a subset of training images. Once the steering committee was satisfied that their scoring procedures were consistent, each steering committee member scored a portion of images (no formal reliability testing was undertaken for this set of readings). All images were read blinded to symptoms, including whether pain was in the hip or knee, and which side was painful. These scores were never made available at subsequent readings.

Independently of steering committee scores, baseline and follow-up images were scored by trained readers and a GP with expertise in OA and radiograph reading (JD) (25). Extensive training was provided to the trained readers (4 readers in years 2 and 5, 5 readers in years 8 and 10) by an experienced musculoskeletal radiologist (EHGO) and the GP, described elsewhere (25). All trained readers were medical students. The GP maintained supervision over trained readers throughout the study, including answering questions or assisting with scores if needed. We previously reported training and interrater reliability using year 5 images (mean prevalence and bias adjusted κ = 0.58 [range 0.23–0.79] for knee K/L scores, and κ = 0.80 [range 0.55–0.90] for the hip)

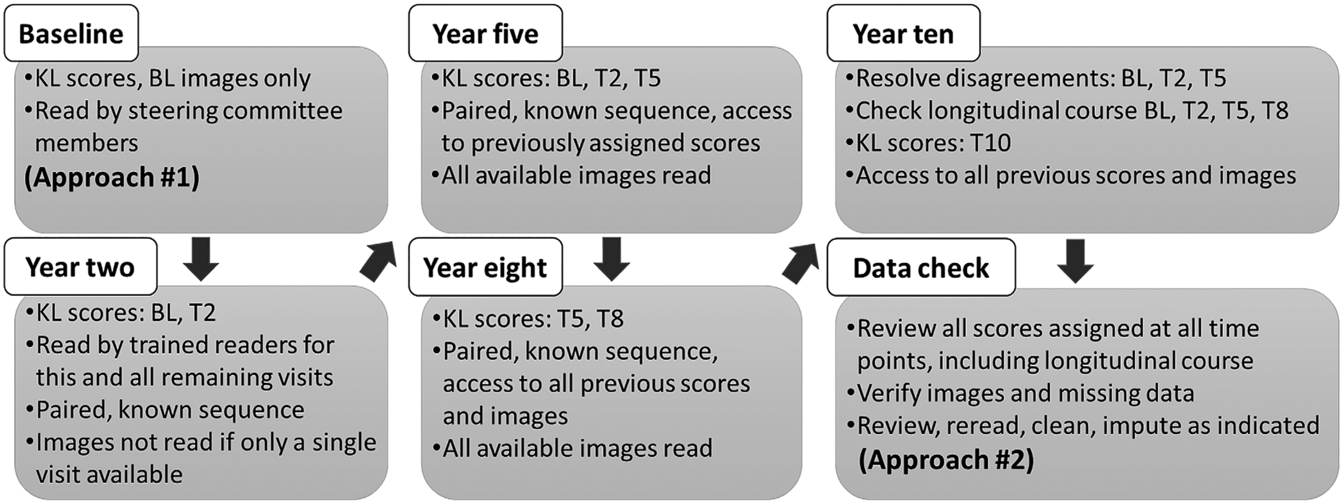
Baseline
•KL scores, BL images only
•Read by steering committee members
(Approach #1)
Year two
•KL scores: BL, T2
•Read by trained readers for this and all remaining visits
•Paired, known sequence
•Images not read if only a single visit available
Year five
•KL scores: BL, T2, T5
•Paired, known sequence, access to previously assigned scores
•All available images read
Year eight
•KL scores: T5, T8
•Paired, known sequence, access to all previous scores and images
•All available images read
Year ten
•Resolve disagreements: BL, T2, T5
•Check longitudinal course BL, T2, T5, T8
•KL scores: T10
•Access to all previous scores and images
Data check
•Review all scores assigned at all time points, including longitudinal course
•Verify images and missing data
•Review, reread, clean, impute as indicated
(Approach #2)

Figure 1. Flow chart of Kellgren/Lawrence (KL) scoring procedures at each visit. BL = baseline; T2, T5, T8, T10 = years 2, 5, 8, 10.

(25). In this article, we differentiate the date of image scoring from the date of image acquisition by spelling out the visit in which images were scored, and abbreviating the visit in which images were acquired: baseline (BL), year 2 (T2), year 5 (T5), year 8 (T8), and year 10 (T10). Thus at baseline, BL images were scored; at year 2, BL and T2 images were scored, and so on.

Baseline. At baseline, scores were assigned by the steering committee without access to follow-up images (Figure 1). For the present study, we defined these K/L scores as the first approach of 2 scoring approaches. The trained readers did not read or score any images at baseline, but began reading images at year 2 (see below). Scores assigned by the steering committee were never made available to trained readers at any time.

Year 2. The trained readers read and scored BL and T2 images paired with known sequence. If T2 images were missing, the BL images were not read or scored. If T2 images were available but the BL image was missing, a K/L score was assigned only to the T2 image at this visit.

Year 5. Trained readers read and scored BL, T2, and T5 radiographs paired with known sequence. Scores were assigned to all available images, even if images were only available for a single visit. Readers had access to previously assigned BL and T2 scores from year 2. Readers could assign different K/L scores for BL and T2 than had previously been assigned, if reading images across all 3 time points together justified this difference.

Year 8. Trained readers read and scored all available T5 and T8 radiographs paired with known sequence, with BL and T2 radiographs available for reference at reader discretion. Readers had knowledge of scores previously assigned at year 5 for BL images. No new scores were assigned for BL or T2.

Year 10. Trained readers first looked at all previous scores that had been read on at least 2 occasions (BL images scored at years 2 and 5; T2 images scored at years 2 and 5; T5 images scored at years 5 and 8). For any case where 2 scores differed, a third read of those images was done to resolve the disagreement, and K/L scores for subsequent time points were checked for longitudinal course. Subsequently, T10 radiographs were read and scored. Readers had access to all previously acquired radiographs and previously assigned scores, but were not explicitly instructed to use them in assigning T10 scores.

Once T10 scoring was complete, all K/L scores across all time points were reviewed with all images available, together in sequence. Further consideration was given to adjusting scores at any time point, if appropriate. For example, in cases where a K/L grade decreased from one time point to the next, all images for that participant were reviewed, and K/L scores were adjusted to better represent images across all time points. This process was done on the assumption that OA cannot regress, thus K/L grades suggesting regression were likely due to variability such as data entry error, interrater error, image quality, or radioanatomic positioning. Other reasons for image rereads included suspected data entry errors or missing scores. At year 10, there were also several cases of images that could not be found from previous time points. We could not confirm whether these images had become missing or if they had never existed, so scores were reassigned to missing. Finally, in cases of missing K/L scores: if the subsequent time point image was K/L 0, then the earlier missing data point was reclassified as K/L 0; if a previous time point had a confirmed joint arthroplasty, then subsequent visits were reclassified to arthroplasty. Remaining missing data were left as missing.

The review of all scores across all time points was done in an iterative manner, with the final review performed in August 2019, including complete score reviews (all scores assigned at all visits) and verification that radiographs existed for all assigned scores (EMM and JR), a team meeting (all coauthors), additional radiograph readings to resolve remaining uncertainties (JD), and approval of the final data set (EMM, JR, and JD). For this study, we defined these final K/L scores as the second approach of 2 scoring approaches.

Joint replacement. Joint replacements were confirmed radiographically. For knees, we defined joint replacement as partial or total arthroplasty. Participants reported the year in which the surgery had occurred, and we recorded this value in years from baseline. If the surgery date was missing, we recorded the date as the visit in which the radiograph of the joint replacement was acquired.

Statistical analysis. All statistical analyses were done using Stata/SE software, version 15.1. We described the proportion of knees or hips with each K/L score (0–4) at baseline using both approaches: the steering committee's single time point reading (first approach), and the trained readers' year-10 final assignment of BL scores with access to images and scores across all time points (second approach). We then reported BL ROA prevalence using both scoring approaches and also using 2 ROA definitions: any ROA (K/L grade ≥1) and established ROA (K/L grade ≥2). We then compared how the 2 scoring approaches affected BL ROA prevalence (any versus established) using mixed-effects Poisson regression with robust estimates of variance.

We next compared the associations of the 4 different BL scores (2 scoring approaches, 2 ROA definitions) with undergoing joint replacement by the end of the study using Cox proportional hazards models (Stata's stcox syntax) (26). To account for correlation between both knees (or hips) within each participant, we clustered models at the participant level using the vce (cluster clustervar) option (26). We defined survival as the year in which a joint replacement occurred, or the year in which participants without joint replacement withdrew, were lost to follow-up, or completed the study.

Finally, we evaluated the associations of the 2 scoring approaches with developing incident established ROA for BL scores of K/L 1 compared to K/L 0 using Cox proportional hazards models. We defined survival as the first visit in which a joint was scored at least K/L 2 (based on the final scores assigned in year 10), or the year in which participants without ROA withdrew, were lost to follow-up, or completed the study.

RESULTS

Of 1,002 participants, 792 (79%) were women, mean ± SD age was 55.9 ± 5.2 years, and body mass index was 26.2 ± 4.0 kg/m^2. BL K/L scores differed between the 2 approaches. Using the first approach, 439 of 1,526 K/L grade 0 knees (29%) were assigned higher scores in the second approach, while 123 K/L grade 1 and 2 scores were assigned K/L 0 in the second approach, resulting overall in 20% fewer K/L 0 scores in the second approach (Table 1 and Figure 2).

Table 1. Kellgren/Lawrence (K/L) scores at baseline in the knee and hip, using 2 scoring approaches: first approach scored by steering committee at baseline without access to follow-up images versus second approach scored by trained readers with all available images and known sequence (n = 2,004 knees)[*]

	First approach	Second approach	PR (95% CI)
Knee K/L score			
0	1,526 (76)	1,228 (61)	–
1	359 (18)	555 (28)	–
2	79 (4)	206 (10)	–
3	8 (<1)	2 (<1)	–
4	0 (0)	0 (0)	–
Missing	32 (2)	13 (<1)	–
Knee radiographic ROA			
Any ROA[†]	446 (22)	763 (38)	1.7 (1.6–1.9)[‡]
Established ROA[§]	87 (4)	208 (10)	2.4 (1.8–3.1)[‡]
Hip K/L score			
0	1,699 (85)	1,292 (64)	–
1	209 (10)	482 (24)	–
2	67 (3)	205 (10)	–
3	7 (<1)	13 (<1)	–
4	0 (0)	0 (0)	–
Missing	22 (1)	12 (<1)	–
Hip radiographic ROA			
Any ROA[†]	283 (14)	700 (35)	2.5 (2.2–2.8)[‡]
Established ROA[§]	74 (4)	218 (11)	2.9 (2.3–3.7)[‡]

[*] Values are the number (%) unless indicated otherwise. 95% CI = 95% confidence interval; PR = prevalence ratio; ROA = radiographic osteoarthritis.
[†] Any ROA = K/L grade ≥1.
[‡] Statistically significant.
[§] Established ROA = K/L grade ≥2.

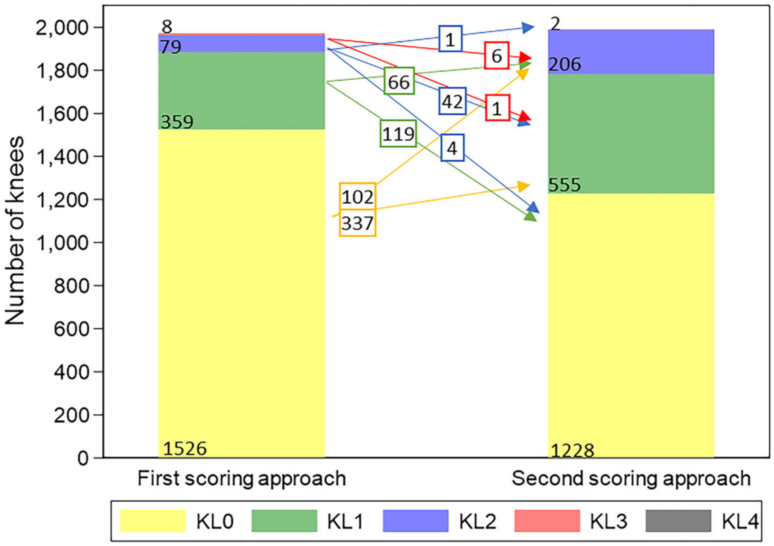

Figure 2. Knee Kellgren/Lawrence (KL) scores: differences in assigned baseline K/L scores by first scoring approach (single reading by steering committee, left column) compared to second scoring approach (paired readings with known sequence, by trained readers, right column). Numbers in columns refer to number of participants assigned each grade; numbers in small boxes refer to number of participants whose grade changed (with arrow indicating to which grade they changed) using the second scoring approach.

Similarly, 485 of 1,699 K/L 0 hips (29%) were assigned higher scores in the second approach, while 69 K/L grade 1 and 2 scores were assigned K/L 0 in the second approach, resulting overall in 24% fewer K/L 0 scores in the second approach (Table 1 and Figure 3).

Using the second approach, more participants were classified as having ROA using both ROA definitions. For knees, the prevalence ratio of the second approach compared to the first was 1.7 (95% confidence interval [95% CI] 1.6–1.9) for any ROA and 2.4 (95% CI 1.8–3.1) for established ROA (Table 1). For hips, prevalence ratios were 2.5 (95% CI 2.2–2.8) and 2.9 (95% CI 2.3–3.7), respectively (Table 1).

The hazard for undergoing knee replacement differed substantially between the 2 scoring approaches, but was only significant for established ROA (Table 2). For any ROA (compared to no ROA) at baseline, the hazard ratio (HR) for undergoing knee replacement was 9.5 (95% CI 4.8–18.6) using the first approach and 13.3 (95% CI 5.4–33.2) using the second

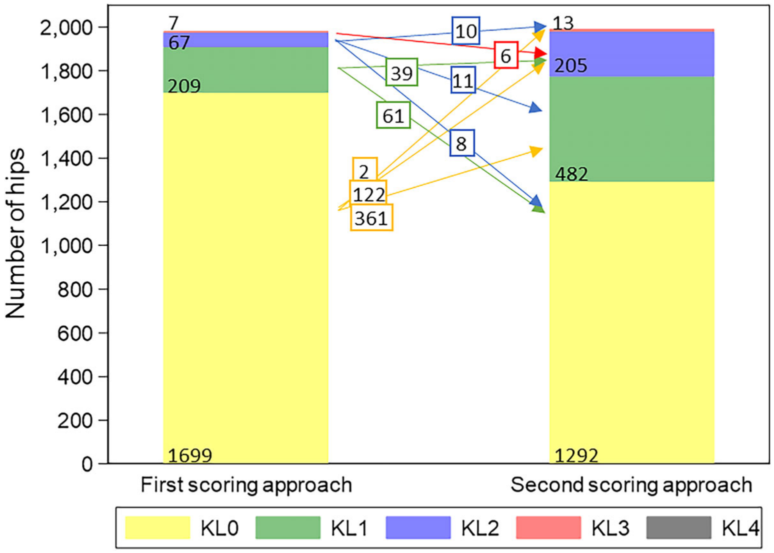

Figure 3. Hip Kellgren/Lawrence (KL) scores: differences in assigned baseline K/L scores by first scoring approach (single reading by steering committee, left column) compared to second scoring approach (paired readings with known sequence, by trained readers, right column). Numbers in columns refer to number of participants assigned each grade; numbers in small boxes refer to number of participants whose grade changed (with arrow indicating to which grade they changed) using the second scoring approach.

Table 2. Hazard ratios for undergoing joint replacement based on baseline for any (K/L grade ≥1) or established (K/L grade ≥2) OA prevalence, using 2 scoring approaches: first approach scored by steering committee at baseline without access to follow-up images versus second approach scored by trained readers with all available images and known sequence*

	First approach	HR (95% CI)	Second approach	HR (95% CI)
Knee replacement				
Any ROA†	32/446 (7)	9.5 (4.8–18.6)‡	39/763 (5)	13.3 (5.4–33.2)‡
K/L 0§	12/1,526 (0.8)	–	5/1,228 (0.4)	–
Established ROA¶	19/87 (22)‡	19.3 (10.3–36.1)‡	15/208 (7)‡	4.8 (2.4–9.6)‡
K/L 0 or 1§	25/1,885 (1)	–	29/1,783 (2)	–
Hip replacement				
Any ROA†	52/283 (18)	9.4 (6.1–14.5)‡	73/700 (10)	8.3 (4.9–14.0)‡
K/L 0§	39/1,699 (2)	–	18/1,292 (1)	–
Established ROA¶	34/74 (46)	24.4 (15.0–39.8)‡	40/218 (18)	7.7 (4.9–12.1)‡
K/L 0 or 1§	57/1,908 (3)	–	51/1,774 (3)	–

* Values are the number/total number (%) unless indicated otherwise. 95% CI = 95% confidence interval; HR = hazard ratio; K/L = Kellgren/Lawrence; OA = osteoarthritis; ROA = radiographic OA.
† Any baseline ROA K/L grade ≥1.
‡ Statistically significant.
§ Without baseline ROA.
¶ Established baseline ROA K/L grade ≥2.

approach. Moreover, 7 more knee replacements (39 of 44 compared to 32) were correctly predicted using the second approach, while at most, 310 knees were reclassified as having any ROA but did not undergo arthroplasty, though on account of right censoring (338 knees [17%]), true results may differ slightly (columns 2 and 4 in Table 2). For established ROA, the HR for undergoing knee replacement was 19.3 (95% CI 10.3–36.1) using the first approach, and decreased significantly to 4.8 (95% CI 2.4–9.6) using the second approach. Using the second approach, 4 fewer arthroplasties were correctly predicted, and at most, 125 more knees with any ROA did not undergo arthroplasty.

For hips, results were similar (Table 2). For any ROA, the HR for undergoing hip replacement was 9.4 (95% CI 6.1–14.5) using the first approach and 8.3 (95% CI 4.9–14.0) using the second approach. Despite similar HRs, the second approach correctly predicted 21 more hip replacements, while up to 396 more hips had any OA but did not undergo arthroplasty. For established ROA, the HR for undergoing hip replacement was 24.4 (95% CI 15.0–39.8) using the first approach and decreased significantly to 7.7 (95% CI 4.9–12.1) using the second approach. Despite the lower HR, the second approach correctly predicted 6 more hip replacements, while at most, 138 more knees with established ROA did not undergo arthroplasty.

The HR for developing incident established knee ROA was 2.4 (95% CI 2.0–2.8) for K/L 1 compared to K/L 0 using the first approach and 2.8 (95% CI 2.4–3.3) using the second approach (Table 3). The second approach correctly predicted 124 more knees developing established ROA, while up to 72 more knees were graded K/L 1 that did not develop ROA. For the hip, the HR was 2.1 (95% CI 1.6–2.7) using the first approach and 3.0 (95% CI 2.5–3.5) using the second approach. The second approach correctly predicted 163 more hips developing established ROA while up to 110 more knees were graded K/L 1 that did not develop ROA.

Table 3. Hazard ratios for developing incident established radiographic OA (K/L grade ≥2) for K/L grade 1 at baseline compared to K/L grade 0, using 2 scoring approaches: first approach scored by steering committee at baseline without access to follow-up images versus second approach scored by trained readers with all available images and known sequence*

	First approach	HR (95% CI)	Second approach	HR (95% CI)
Knee				
K/L 1†	269/359 (75)	2.4 (2.0–2.8)‡	393/555 (71)	2.8 (2.4–3.3)‡
K/L 0§	734/1,526 (48)	–	494/1,228 (40)	–
Hip				
K/L 1†	129/209 (62)	2.1 (1.6–2.7)‡	292/482 (61)	3.0 (2.5–3.5)‡
K/L 0§	706/1,699 (42)	–	396/1,292 (31)	–

* Values are the number/total number (%) unless indicated otherwise. 95% CI = 95% confidence interval; HR = hazard ratio; K/L = Kellgren/Lawrence; OA = osteoarthritis.
† Incident established OA with baseline K/L grade 1.
‡ Statistically significant.
§ Incident established radiographic OA with baseline K/L grade 0.

DISCUSSION

In this study, we described the methods used in the CHECK cohort to assign K/L scores to hip and knee radiographs at each visit. With these details consolidated into a single article, the reader is better equipped to compare and interpret studies published since the CHECK cohort's inception, that use K/L scores assigned at different time points. This study also illustrates how different scoring methods potentially influence cohort study results, highlighting potential implications for future trial design and interpretation.

The second scoring approach classified more hips and knees as having both any and established ROA compared to the first approach. This difference may be due to inherent challenges in determining whether a bony feature is an osteophyte, and whether it is doubtful or definite. Seeing follow-up images with progression of osteophytes may increase reader confidence in identifying and classifying baseline features as osteophytes. We acknowledge, however, that this difference could also relate to who assigned scores under the 2 approaches. Interrater reliability has previously been shown to be higher between expert radiologists than between expert radiologists and their trained readers (17,27,28). We therefore acknowledge that the differences between the 2 approaches in our study may reflect not only access to follow-up images, but also interrater reliability and relative expertise and training of the 2 groups of readers. One previous study reported that, among disagreements between an expert radiologist and trained readers, scores tended to be higher in trained readers (17). These findings are similar to ours. However, readers in that study were site investigators motivated to enroll individuals with OA features into their study, possibly introducing bias (17). Our study eligibility criteria did not include radiograph readings, removing this bias. We believe that higher scores in the second approach are more likely due to access to follow-up images and extensive data checking, though we cannot rule out reader-related factors.

Previous studies have shown that reading images paired in known sequence improves reliability and sensitivity to ROA progression, likely due to having access to more information during reading (4–11). Sensitivity to progression has been implied to suggest that, despite the bias introduced, paired reading with known sequence provides more valid scores. However, sensitivity to progression has typically been defined using the standardized response mean (SRM) (4–6). This statistic provides the equivalent of a mean effect size, so a larger SRM means more individuals are reported to have ROA progression. A gold standard has not typically been considered to confirm that larger SRMs reflect a true higher rate of progression (29). Thus while this approach may be more valid, SRM cannot confirm this increase in validity. At best, SRM provides face validity that having access to more images enables a more accurate score, but we cannot rule out that a higher SRM reflects bias introduced by a reader expecting progression to occur chronologically. Reading an image at a single time point may increase error and reduce reliability. However, such a reading also mitigates bias, may give more conservative estimates of ROA prevalence, and better reflects clinical settings where multiple images are not available.

One of the strengths of the CHECK cohort is that 2 approaches have been used to assign baseline K/L scores. This offers the unique ability to select which approach would answer specific questions best. For example, if a researcher wants to know whether baseline K/L scores are a risk factor for a future outcome, they could use scores assigned using the second approach because this method is more accurate (29). Alternatively, if researchers want to know how well radiographs in a clinical setting predict the same outcome, the first approach may provide a more conservative and clinically realistic estimate, since clinicians do not typically know the outcomes of care provided.

Our results highlight the importance of reporting absolute numbers of an outcome in addition to effect sizes: odds ratios, relative risks, or HRs reported alone may be misleading. For example, if a clinician wants to identify hip pain patients at risk for future hip replacement to be able to offer a cost-effective prevention program, the clinician could use the results of the more clinically realistic first approach (Table 2). They might be tempted to define ROA as K/L ≥2 because of the higher HR (24.4 compared to 9.4 if defining ROA as K/L ≥1). However, defining ROA as K/L ≥2 would result in not treating 18 hips that would need a future hip replacement and may benefit from treatment. In this case, treating any hip ROA despite the lower HR might be more important to the clinician. In the case of an expensive treatment, the clinician might stay with K/L 2 after all because while they would miss treating the 18 hips ultimately needing replacement, in this scenario, a clinician would theoretically avoid the need to provide costly treatment for more than 200 hips. This scenario also illustrates that the number of patients the clinician might expect to treat would be substantially overestimated had they implemented a new program based on results using the second scoring approach (700 knees with K/L 1, 218 knees with K/L 2).

The above scenario brings up the additional question of how best to define ROA. In a previous 10-year prospective population-based study of women, 62% of 90 knees with doubtful osteophytes at baseline progressed to having definite osteophytes 10 years later (18). Our findings were similar: 71–75% of knees (depending on scoring approach) and 61–62% of hips with doubtful osteophytes at baseline developed established ROA within 10 years. These results suggest that identifying middle-aged individuals with hip or knee symptoms as having OA, rather than waiting for them to develop established ROA, may offer new insights and opportunities for secondary prevention in this population.

Limitations to our study include the fact that interrater reliability was not formally assessed in the steering committee of expert readers, and trained reader reliability was assessed at years 5 and 8 but only recorded at year 5. All readers were of similar background and received similar training by the same radiologist and GP, thus the recording of year 8 results was not felt to be necessary at the time. Also, to more accurately compare scoring methods, having the same readers assign scores using both approaches would have been advantageous. The comparisons of scoring approaches in our study represent a more pragmatic and thus generalizable comparison of approaches that capture differences due in part to having access to multiple follow-up images in known sequence, but also due to interrater reliability, differences in data-checking procedures, and reader-related factors. In addition, all participants had knee or hip symptoms, thus we had no asymptomatic reference group. However, our study design better reflects clinical reality in which patients typically seek care for existing symptoms. Finally, in the CHECK cohort, the reason for study withdrawal was not recorded. This limitation relates to use of survival analysis, and the possibility of competing risk, in particular death. While we cannot confirm this fact, the young age of our participants, combined with recollection of study investigators, suggests that death was very rare and our findings would not likely be altered.

We recommend that future studies be designed with careful consideration for radiographic scoring conditions. Evaluating a score assigned with a single reading at a single visit may give more clinically realistic predictions of future outcomes. Alternatively, evaluating scores assigned during paired readings with a known sequence may provide greater insights into the exact nature of ROA onset and progression. Both approaches are important, and thus method selection must address the specific research question. In both cases, readers must be carefully selected, with adequate experience and training to optimize score validity and reliability. We also recommend that future studies consider using an earlier definition of ROA than is typically used, particularly where researchers are interested in understanding early OA with an aim toward preventing poor clinical outcomes. Where feasible and affordable, studies incorporating magnetic resonance imaging (MRI) can also contribute meaningfully to early OA research, since MRI better visualizes soft tissues (e.g., cartilage, bone marrow lesions) and is thus more sensitive to detecting early OA features (30). In conclusion, this study of middle-aged individuals with hip or knee symptoms demonstrates that evaluation of ROA depends on radiograph scoring conditions, and the prediction of future outcomes is influenced by both scoring conditions as well as which K/L grade is used to define ROA.

ACKNOWLEDGMENTS

The CHECK cohort study was initiated by the Dutch Arthritis Association, and participating sites were: Erasmus Medical Center Rotterdam; Kennemer Gasthuis Haarlem; Leiden University Medical Center; Maastricht University Medical Center; Martini Hospital Groningen/Allied Health Care Center for Rheumatology and Rehabilitation Groningen; Medical Spectrum Twente Enschede/Ziekenhuisgroep Twente Almelo; Reade (formerly Jan van Breemen Institute/VU Medical Center Amsterdam); St. Maartenskliniek Nijmegen; and University Medical Center Utrecht and Wilhelmina Hospital Assen.

AUTHOR CONTRIBUTIONS

All authors were involved in drafting the article or revising it critically for important intellectual content, and all authors approved the final version to be submitted for publication. Dr. Macri had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Study conception and design. Macri, Runhaar, Damen, Oei, Bierma-Zeinstra.

Acquisition of data. Runhaar, Damen, Oei, Bierma-Zeinstra.

Analysis and interpretation of data. Macri, Runhaar, Damen, Oei, Bierma-Zeinstra.

REFERENCES

1. Kellgren JH, Lawrence JS. Radiological assessment of osteo-arthrosis. Ann Rheum Dis 1957;16:494–502.
2. Schiphof D, Boers M, Bierma-Zeinstra SM. Differences in descriptions of Kellgren and Lawrence grades of knee osteoarthritis. Ann Rheum Dis 2008;67:1034–6.
3. Kellgren J, Jeffrey M, Ball J. The epidemiology of chronic rheumatism: atlas of standard radiographs of arthritis. Oxford (UK): Blackwell Scientific Publications; 1963.
4. Auleley GR, Giraudeau B, Dougados M, Ravaud P. Radiographic assessment of hip osteoarthritis progression: impact of reading procedures for longitudinal studies. Ann Rheum Dis 2000;59:422–7.
5. Gensburger D, Roux JP, Arlot M, Sornay-Rendu E, Ravaud P, Chapurlat R. Influence of blinding sequence of radiographs on the reproducibility and sensitivity to change of joint space width measurement in knee osteoarthritis. Arthritis Care Res (Hoboken) 2010;62: 1699–705.
6. Botha-Scheepers S, Watt I, Breedveld FC, Kloppenburg M. Reading radiographs in pairs or in chronological order influences radiological progression in osteoarthritis. Rheumatology (Oxford) 2005;44: 1452–5.
7. Van der Heijde D, Boonen A, Boers M, Kostense P, van Der Linden S. Reading radiographs in chronological order, in pairs or as single films has important implications for the discriminative power of rheumatoid arthritis clinical trials. Rheumatology (Oxford) 1999;38:1213–20.
8. Felson DT, Nevitt MC, Yang M, Clancy M, Niu J, Torner JC, et al. A new approach yields high rates of radiographic progression in knee osteoarthritis. J Rheumatol 2008;35:2047–54.
9. LaValley MP, McAlindon TE, Chaisson CE, Levy D, Felson DT. The validity of different definitions of radiographic worsening for longitudinal studies of knee osteoarthritis. J Clin Epidemiol 2001;54:30–9.
10. Kopec JA, Sayre EC, Schwartz TA, Renner JB, Helmick CG, Badley EM, et al. Occurrence of radiographic osteoarthritis of the knee and hip among African Americans and Whites: a population-based prospective cohort study. Arthritis Care Res (Hoboken) 2013;65: 928–35.
11. Spector TD, Hart DJ, Byrne J, Harris PA, Dacre JE, Doyle DV. Definition of osteoarthritis of the knee for epidemiological studies. Ann Rheum Dis 1993;52:790–4.
12. The Osteoarthritis Initiative. Central reading of knee X-rays for Kellgren & Lawrence grade and individual radiographic features of tibiofemoral knee OA. URL: http://oai.epi-ucsf.org/datarelease/forms/kXR_SQ_BU_Descrip.pdf?V01XRKL.
13. Schiphof D, de Klerk BM, Kerkhof HJ, Hofman A, Koes BW, Boers M, et al. Impact of different descriptions of the Kellgren and Lawrence classification criteria on the diagnosis of knee osteoarthritis. Ann Rheum Dis 2011;70:1422–7.
14. Chaisson CE, Gale DR, Gale E, Kazis L, Skinner K, Felson DT. Detecting radiographic knee osteoarthritis: what combination of views is optimal? Rheumatology (Oxford) 2000;39:1218–21.

15. Radiography Working Group of the OARSI-OMERACT Imaging Workshop, Le Graverand MP, Mazzuca S, Lassere M, Guermazi A, Pickering E, et al. Assessment of the radioanatomic positioning of the osteoarthritic knee in serial radiographs: comparison of three acquisition techniques. Osteoarthritis Cartilage 2006;14:37–43.
16. Portney LG, Watkins MP. Foundations of clinical research: applications to practice. Third ed. Upper Saddle River (NJ): Pearson Eduction; 2009.
17. Guermazi A, Hunter DJ, Li L, Benichou O, Eckstein F, Kwoh CK, et al. Different thresholds for detecting osteophytes and joint space narrowing exist between the site investigators and the centralized reader in a multicenter knee osteoarthritis study: data from the Osteoarthritis Initiative. Skeletal Radiol 2012;41:179–86.
18. Duncan R, Peat G, Thomas E, Hay EM, Croft P. Incidence, progression and sequence of development of radiographic knee osteoarthritis in a symptomatic population. Ann Rheum Dis 2011;70:1944–8.
19. Hart DJ, Spector TD. Kellgren & Lawrence grade 1 osteophytes in the knee: doubtful or definite? Osteoarthritis Cartilage 2002;11:149–50.
20. Wesseling J, Boers M, Viergever MA, Hilberdink WK, Lafeber FP, Dekker J, et al. Cohort profile: cohort hip and cohort knee (CHECK) study. Int J Epidemiol 2014;45:36–44.
21. Schiphof D, Runhaar J, Waarsing JH, van Spil WE, van Middelkoop M, Bierma-Zeinstra SM. The clinical and radiographic course of early knee and hip osteoarthritis over 10 years in CHECK (Cohort Hip and Cohort Knee). Osteoarthritis Cartilage 2019;27:1491–500.
22. Damen J, van Rijn RM, Emans PJ, Hilberdink WK, Wesseling J, Oei EH, et al. Prevalence and development of hip and knee osteoarthritis according to American College of Rheumatology criteria in the CHECK cohort. Arthritis Res Ther 2019;21:4.
23. Lankhorst N, Damen J, Oei E, Verhaar J, Kloppenburg M, Bierma-Zeinstra S, et al. Incidence, prevalence, natural course and prognosis of patellofemoral osteoarthritis: the Cohort Hip and Cohort Knee study. Osteoarthritis Cartilage 2017;25:647–53.
24. Van Spil WE, Welsing PM, Kloppenburg M, Bierma-Zeinstra SM, Bijlsma JW, Mastbergen SC, et al. Cross-sectional and predictive associations between plasma adipokines and radiographic signs of early-stage knee osteoarthritis: data from CHECK. Osteoarthritis Cartilage 2012;20:1278–85.
25. Damen J, Schiphof D, Ten Wolde S, Cats HA, Bierma-Zeinstra SM, Oei EH. Inter-observer reliability for radiographic assessment of early osteoarthritis features: the CHECK (Cohort Hip and Cohort Knee) study. Osteoarthritis Cartilage 2014;22:969–74.
26. Stata Corporation. Stata survival analysis reference manual release 16. College Station (TX): Stata Press; 2019.
27. Bellamy N, Tesar P, Walker D, Klestov A, Muirden K, Kuhnert P, et al. Perceptual variation in grading hand, hip and knee radiographs: observations based on an Australian twin registry study of osteoarthritis. Ann Rheum Dis 1999;58:766–9.
28. Cooper C, Cushnaghan J, Kirwan J, Dieppe P, Rogers J, McAlindon T, et al. Radiographic assessment of the knee joint in osteoarthritis. Ann Rheum Dis 1992;51:80–2.
29. Felson DT, Nevitt MC. Blinding images to sequence in osteoarthritis: evidence from other diseases. Osteoarthritis Cartilage 2009;17: 281–3.
30. Gudbergsen H, Lohmander L, Jones G, Christensen R, Bartels EM, Danneskiold-Samsoe B, et al. Correlations between radiographic assessments and MRI features of knee osteoarthritis: a cross-sectional study. Osteoarthritis Cartilage 2013;21:535–43.